

# Coarse-grained Cross-lingual Alignment of Comparable Texts with Topic Models and Encyclopedic Knowledge

**Vivi Nastase**  
HLT group  
FBK, Trento, Italy  
nastase@fbk.eu

**Angela Fahrni**  
HITS gGmbH  
Heidelberg, Germany  
angela.fahrni@h-its.org

## Abstract

We present a method for coarse-grained cross-lingual alignment of comparable texts: segments consisting of contiguous paragraphs that discuss the same theme (e.g. history, economy) are aligned based on induced multilingual topics. The method combines three ideas: a two level LDA model that filters out words that do not convey themes, an HMM that models the ordering of themes in the collection of documents, and language-independent concept annotations to serve as a cross-language bridge and to strengthen the connection between paragraphs in the same segment through concept relations. The method is evaluated on English and French data previously used for monolingual alignment. The results show state-of-the-art performance in both monolingual and cross-lingual settings.

## 1 Introduction

Coarse-grained alignment of documents groups text segments in different documents that convey the same theme – e.g. *history*, *geography* in texts about cities. Cross-language alignment deals with the added challenge of aligning segments from documents in different languages. This could be useful as a prelude to fine-grained alignment, or for building coarsely aligned multilingual corpora for machine translation or text categorization. We are interested in cross-lingual alignment for the synchronization of web (wiki) pages with multiple language versions, where pages in different languages are independently edited. A coarse alignment would (i) reveal quickly text portions that are not shared, and must thus be translated and added, and (ii) show potentially parallel portions, to be further processed to produce a more fine-grained alignment. We present a cross-lingual

alignment of segments – consisting of one or more paragraphs – based on knowledge-enhanced topic modeling, illustrated in Figure 1, which implements the following ideas:

**2-level LDA** A two-layer modeling framework (i) models paragraphs as a mixture of three topics – background, document-specific and theme-specific (e.g. in a document about the city of Montreal, articles, prepositions, etc. would be background, words associated with Montreal but not specific to a theme – Montreal, Canada, French, Quebec – would be document-specific, and others such as hockey, sports, skating would be theme-specific); (ii) models a document as a mixture of  $K$  topics/themes.

**HMM theme sequences** Following the assumption that themes are not presented randomly in a document, we use an HMM to model their sequence. State transitions are modeled through a Dirichlet distribution, tuned to bias the system towards self-transitions and thus avoid rapid switching between states (Beal et al., 2002; Teh et al., 2003; Fox et al., 2010).

### Language-independent concept annotations

We inject knowledge in the model by linking single and multi-word terms to concepts in a concept network obtained from Wikipedia (Nastase and Strube, 2013). Terms are replaced with language-independent identifiers (e.g. “New York” becomes “c553795”). Enriching text with concept annotations allows to: (i) integrate naturally the identified multi-word expressions into the statistical process; (ii) deal with ambiguity (“New York” the city becomes “c553795”, while “New York” the state becomes “c27491610”); (iii) deal partly with coreferent mentions and synonymy (“New York City”, “The Big Apple”, “Nueva York” share the same identifier); (iv) bridge languages and build multilingual topic models, as concepts are shared across languages;

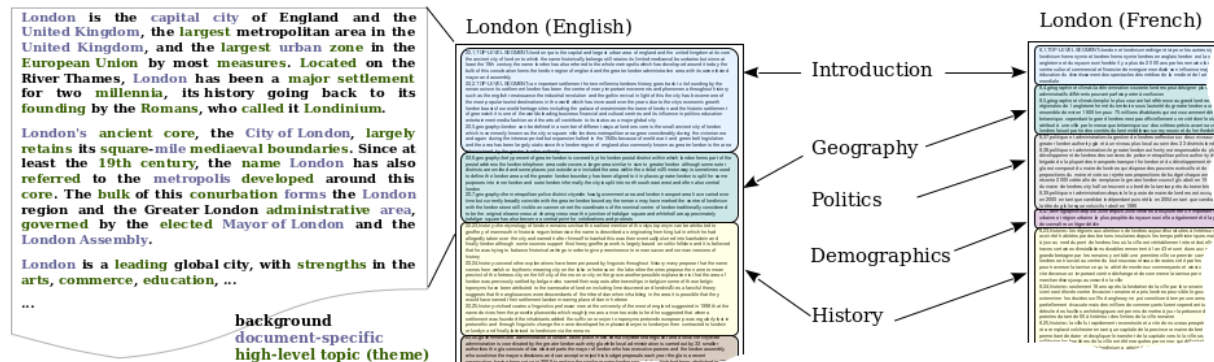


Figure 1: A two level – paragraph and document – topic modeling for cross-document and cross-lingual topic-based alignment. Within a paragraph, words pertaining to different topics are highlighted with different colours.

(v) use additional encyclopedic knowledge extracted from the concept network, e.g. relations between concepts within the same or consecutive paragraphs to strengthen the cohesion of the topic.

The method is applied on Wikipedia articles about cities in English and French (Chen et al., 2009). This allows for comparison with related work (in the monolingual setting), and to extend the evaluation to a cross-lingual setting. The monolingual segment alignment is evaluated against a Hidden Topic Markov Model (Gruber et al., 2007) – and the GMM model (Chen et al., 2009). The method described here scores higher than both. In the cross-lingual setting, the baselines are the alignments produced using translation tables, and using concept annotations, respectively. The method which combines topic modeling with concept annotations outperforms both baselines.

## 2 Previous Work

**Text alignment** For monolingual comparable corpora, text relatedness measures produce good alignment results (Barzilay and Elhadad, 2003; Yahyaei et al., 2011). Methods for sentence alignment in a parallel bilingual corpus typically rely on linearity of the alignment, existence of 1:1 mapping between aligned sentences and the correlation of sentence length (Tiedemann, 2011). These assumptions do not hold for comparable (but not parallel) corpora. Bilingual comparable corpora have been mainly used to extract parallel fragments (Gupta et al., 2013) or paraphrases and word translations (Fung and McKeown, 1997). We focus on an alignment of fragments that have the same theme, but are not necessarily parallel.

**Sequence modeling** Modeling a document as a sequence of topics brings HMMs naturally to mind. Work in this area started with Mulbregt et al. (1998) and Blei and Moreno (2001). This approach suffers from two shortcomings: the number of states needs to be fixed (to be able to perform the Baum-Welch or Viterbi algorithms), and the HMM tends to switch fast between different states. In situations where a state should be persistent, as in topic segmentation, this is a problem. A breakthrough has come when state transitions in an HMM were modeled through a Dirichlet process (Beal et al., 2002). This allows one to control state transitions, and to let the model induce the number of states that best fit the data. Beal et al. (2002) used three hyperparameters to control the HMM: for self-transitions, for transitions to previously used states, and for adding a new transition. Teh et al. (2003) and Fox et al. (2010) extended this model and showed how to manipulate parameters to avoid rapid switching between states.

**Language models** In topic segmentation one can consider the words to be the basic unit (Gruber et al., 2007), or sentences/paragraphs (e.g. (Blei and Moreno, 2001), (Eisenstein and Barzilay, 2008)). Sentences are themselves heterogeneous. Daumé III and Marcu (2006) used topic models for query-driven summarization. The assumption is that each sentence consists of a mixture of language models – one corresponding to the general model of the English language, one corresponding to the overall theme of the document, and one that matches the topic of the given query. Zhai et al. (2004), Titov and McDonald (2008) and Paul and Girju (2009) model topics (aspects) that run throughout a collection of documents, and

topics that are collection-specific. Considering a sentence/paragraph as a small document, and one document as a collection of these small documents, it could be construed that the previously mentioned work models a sentence/paragraph as a mixture of topics, some of which are common throughout the document, some of which are sentence-specific. Wang et al. (2011) also assume a sentence to be generated from a mixture of topics, in particular two: a “functional” (i.e. background) topic, and a content topic. Additionally, this approach models transitions between topics through an HMM, but like previous work, without concerns about the rapid state switching. This does not pose a problem because of the nature of the data: short ads and reviews with short sentences and not much topic repetition.

**WSD and topic modeling** Boyd-Graber et al. (2007) combine topic modeling with word sense disambiguation relative to WordNet, following the observation that words in different topics may exhibit different senses, and that the sense of a word depends on the context in which it appears. Each noun that appears in WordNet will enhance the probabilities of all paths from the root to each possible sense, and “correct” paths will have higher probability because of aggregated evidence from the words in the text. A downside of this is that, as topic models do, it deals with single word terms only, whereas texts contain numerous multi-word terms referring to real world entities. For this reason we apply the concept and entity identification process before the topic modeling step.

**Multilingual topic modeling** Jagarlamudi and Daumé III (2010) use a bilingual dictionary to obtain multilingual topics from an unaligned multilingual corpus. They assume topics to be formed of concepts, which can have different lexicalizations depending on the language. These concepts consist of entries from a bilingual dictionary. Boyd-Graber and Blei (2009) simultaneously discover matching words and multilingual topics from a collection composed of documents in two languages, based on the assumption that similar words appear in similar contexts. Zhang et al. (2010) use a bilingual dictionary as a source of constraints for bridging texts in two languages, using the assumption that related words in different languages have similar distributions. The topic models produced contain words in different languages that are allowed to have different probabil-

ities, reflecting the difference between the data in the two languages. Ni et al. (2009) mine multilingual topics from Wikipedia, using the articles on the same theme in different languages as a source of language models. Mimno et al. (2009) build polylingual topic models from sets of documents on the same topics, in particular Wikipedia articles in several languages. They assume that the different language versions of an article have similar topic distributions, and topics consist of language-specific word distribution. Vulić et al. (2011) build bilingual topic models for information retrieval, using comparable corpora – in particular collections of Wikipedia articles on the same topics in the targeted languages.

### 3 A Topic Model for Alignment

We build upon some of the ideas presented above to develop a knowledge-enhanced Hidden Topic Markov Model (HTMM): documents, made up of paragraphs, are modeled as sequences of topics, with transition between states controlled by a Dirichlet distribution. The process is detailed below.

#### 3.1 Overall generative process

We describe here the overall generative process, covering the two level LDA and HMM modeling, which will be detailed separately below.

Assuming  $K$   $t$ -topics (themes) represented in the documents  $d_{1:M}$  in our collection, the generative process is as follows:

- 1 Draw a word distribution  $\xi_1 \sim \text{Dirichlet}(\eta_1)$  for a background language model;
- 2 Draw a word distribution  $\phi_z \sim \text{Dirichlet}(\beta)$  for each  $t$ -topic  $z \in \{1..K\}$ ;
- 3 draw a  $t$ -topic transition probability distribution  $\pi \sim \text{Dirichlet}(\alpha + \kappa)$  ( $\pi_0$  is an initial state probability vector,  $\pi_j$  is a transition probability distribution from state  $j$ );
- 4 For each document  $d_m, m = 1..M$ :
  - 4.1 draw a word distribution  $\xi_{2m} \sim \text{Dirichlet}(\eta_2)$  for a document-specific language model;
  - 4.2 draw a  $t$ -topic mixture  $\theta_m \sim \text{Dirichlet}(\lambda)$  for document  $d_m$ ;
  - 4.3 for each paragraph (sequence of words)  $\mathbf{w}_{tm}$  in  $d_m$ :
    - 4.3.1 sample a  $t$ -topic  $z_t \sim \text{Discrete}(\pi_{z_{t-1}}, \theta_m)$ ;
    - 4.3.2 draw a  $w$ -topic mixture  $\psi_m \sim \text{Dirichlet}(\gamma)$ ;
    - 4.3.3 for each word  $w_{ti}$  in sequence  $\mathbf{w}_{tm}$ :
      - A. sample a topic  $s_{ti} \sim \text{Discrete}(\psi_m)$ ;
      - B. if  $s_{ti} = 1$ , sample  $w_{ti}$  from the background language model  $\xi_1$ ;
      - C. if  $s_{ti} = 2$ , sample  $w_{ti}$  from the document-specific language model  $\xi_{2m}$ ;
      - D. if  $s_{ti} = 3$ , sample  $w_{ti}$  from the  $t$ -topic  $\phi_{z_t}$ .

[...] in 2006 montreal was named a unesco city of design only one of three cities [...]  
 [...] in 2006 c7954681 was named a c21786641 c8560 of c2318702 only one of three c8560 [...]  
 [...] the biggest sport following in montreal clearly belongs to hockey [...]  
 [...] the biggest c25778403 following in c7954681 clearly belongs to c10886 [...]

Figure 2: Text fragments before and after concept identification.

### 3.2 A 2-level LDA

**Paragraphs** are modeled as generated from a mixture of three *word(-level) topics* (*w-topics*) corresponding to each of (1) background, (2) document-specific, (3) “theme-specific” language model respectively, exemplified in the left side of Figure 1. This modeling is used as a filter – the (collection-specific) background and the document-specific words are not informative with respect to themes, and are skipped in the next processing step. While only side-effects with respect to the focus of this paper, the background and document-specific language models can be useful: the document-specific word probability distribution can be used to align, cluster or classify documents. We explore this briefly in Section 4.1.2. The *w-topic* mixture follows a Dirichlet distribution with hyperparameter  $\gamma$  (*Dirichlet*( $\gamma$ )), while words within a *w-topic* follow *Dirichlet*( $\eta$ ). The topic  $s_i$  assigned to word at position  $i$  in paragraph  $t$  is sampled according to:

$$p(s_i = l | \mathbf{s}, \mathbf{w}_t) \propto g_l(w_{ti}) \frac{n_{w_{ti}}^l + \eta}{n_*^l + W\eta} \frac{n_{t-i}^l + \gamma}{n_{t-i}^* + 3\gamma}$$

where  $W$  is the size of the vocabulary,  $n_w^l$  is the number of times word  $w$  was assigned to *w-topic*  $l$ ,  $*$  is a wild-card,  $n_t^l$  is the number of words in paragraph  $t$  assigned to topic  $l$ ,  $n_t^*$  is the number of words in paragraph  $t$ , and  $\mathbf{w}_{t-i}$  is the sequence of words  $\mathbf{w}_t$  minus the word  $w_{ti}$  at position  $i$ .  $g_{1..3}$  are normalized coefficients that capture the bias of each word for the three postulated topics, according to their document and collection frequencies<sup>1</sup>.

<sup>1</sup>Background language  $g_1(w_{ti}) = \frac{c_{w_{ti}}^P}{c_*^P}$  (generally frequent words); document-specific  $g_2(w_{ti}) = \frac{c_{w_{ti}}^d}{c_*^d}$  (common throughout the document, rare outside); topic-carrying  $g_3(w_{ti}) = \frac{c_{w_{ti}}^D}{c_*^D} \times (1 - g_1(w_{ti}))$  (appear throughout the collection, but not frequent within single documents).  $P$  is the total number of paragraphs in the set of documents  $D$ ,  $d$  is a document (the document being processed),  $c_x^y$  is the number of  $y$ s in which  $x$  appears. Because  $g_l$  are constant, they will not be affected by marginalizing  $\xi$ :  $p(\mathbf{w}_t, \mathbf{g} | \mathbf{s}, \eta) = \int p(\mathbf{w}_t, \mathbf{g} | \mathbf{s}, \xi) p(\xi | \eta) d\xi = \int p(\mathbf{g} | \mathbf{s}) p(\mathbf{w}_t | \mathbf{s}, \xi) p(\xi | \eta) d\xi = \prod_i g_{s_i}(w_{ti}) \int p(\mathbf{w}_t | \mathbf{s}, \xi) p(\xi | \eta) d\xi$ . These factors are normalized such that for each word they sum to 1.

**Documents** are generated from a mixture of  $K$  *themes* (*t-topics*) following *Dirichlet*( $\lambda$ ). The *t-topic* of a paragraph is determined based on the subset of its *theme-specific words*, as determined by the previous step. The words in a topic follow *Dirichlet*( $\beta$ ). The right side of Figure 1 shows an example of the *t-topics* within a document.<sup>2</sup> Contiguous paragraphs that express the same *t-topic* form a segment, and through this can be aligned across documents and languages. Paragraph *t-topics* are sampled as:

$$p(z_t = j | \mathbf{z}, \mathbf{w}_t, \lambda, \beta) \propto \frac{n_{w_{ti}}^j + \beta}{n_*^j + W\beta} \frac{n_{t-i}^j + \lambda}{n_{t-i}^* + K\lambda}$$

### 3.3 Modeling the sequence of topics/themes

The sequence of themes within a document is modeled through an HMM. The transition probabilities are modeled by a Dirichlet distribution (Beal et al., 2002): the transition probabilities from a state  $z_t = j$  at time  $t$  can be interpreted as mixing proportions for the state at time  $t + 1$ :  $\pi_j = \{\pi_{j1}, \dots, \pi_{jK}\}$ . The **persistence of states** is encouraged by increasing the probability of self-transitions using a “sticky” parameter  $\kappa$  (Beal et al., 2002; Fox et al., 2010):  $\pi_i | \alpha, \kappa \sim \text{Dirichlet}(\alpha + \kappa \delta_i)$ , where  $\kappa > 0$  is added to the  $i^{\text{th}}$  component of the parameter vector  $\alpha$ <sup>3</sup>. *t-topics* are sampled as:

$$p(z_t = j | z_{-t}, \alpha, \kappa, \lambda) \propto$$

$\frac{n_{-t,j}^{(d)} + \lambda}{\sum_l n_{-t,l}^{(d)} + K\lambda}$	topic mixing proportion
$\frac{n_{z_{t-1}j}^{-t} + \alpha_j + \kappa * \delta(z_{t-1}, j)}{n_{z_{t-1}*}^{-t} + \sum_x \alpha_x + \kappa}$	transition from previous paragraph
$\frac{n_{jz_{t+1}}^{-t} + \alpha_j + \kappa * \delta(j, z_{t+1})}{n_{j*}^{-t} + \sum_x \alpha_x + \kappa}$	transition to next paragraph

where  $\delta(i, j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$  (Kronecker’s delta).

<sup>2</sup> The *t-topics* induced by the model are nameless, we name them in the figure to illustrate more clearly the point.

<sup>3</sup> Fox et al. (2010) model the transitions through a Dirichlet process to allow the model to infer the number of states.

After sampling the  $t$ -topic for paragraph  $t$ , the  $t$ -topics of the topic-carrying words in  $\mathbf{w}_t$  (assigned  $w$ -topic 3) are reassigned to the paragraph's  $t$ -topic, reflecting the assumption that there is one  $t$ -topic per paragraph.

### 3.4 Language-independent concepts

To bridge languages, and address (at least partly) the issues of multi-word expressions, synonymy, polysemy, we introduce concept annotations, exemplified in Figure 2. To identify concepts:

- locate possible concept lexicalizations extracted from Wikipedia in texts;
- for each lexicalization, find all candidate concepts it could refer to;
- disambiguate among possible candidate concepts.

For disambiguation, the text is represented as a complete  $n$ -partite graph  $G = (V_1, ..V_n, E)$ . Each partition  $V_i$  corresponds to a (possibly multi-word) term  $t_i$  in the text, and contains as vertices all concepts  $c_{ij}$  that may be expressed by  $t_i$ , found in the first step. Each vertex from a partition is connected to all vertices from the other partitions (making the  $n$ -partite graph complete) through edges  $e_{v_i, v_j} \in E$  weighted by  $w_{v_i, v_j}$ . The weights are learned from Wikipedia's link structure, using as features several measures: shared outgoing links between the articles corresponding to the two concepts, shared categories and their specificity; preference of each concept for the other concept's expression through the anchor in the current text.

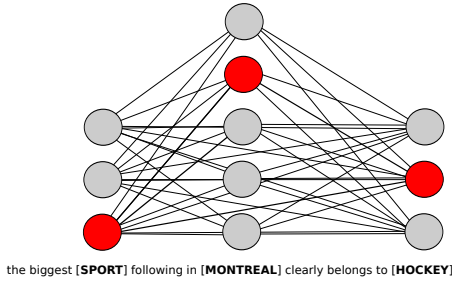


Figure 3: Weighted  $n$ -partite graph. Edge weights are represented by thickness.

The disambiguated concepts are the nodes in the maximum edge-weighted clique in this graph, and their expression in the text is replaced with the corresponding unique (and language independent) concept ID, resulting in the concept-annotated texts illustrated in Figure 2 (Fahrni et al., 2011).

This concept identification method scored high on both ACE 2005 (72.7% F-score) and in the Entity Linking TAC 2011 task.

The texts with concept identifiers are processed with the previously described topic model. To take advantage of the relations between concepts, we introduce a factor  $s_{wkn}$ , to give a boost to topics that are favoured by the concepts  $w_r \in \mathcal{R}_{w_{ti}}$  directly connected to  $w_{ti}$  in the concept network (for  $w_{ti}$  that represent a concept):

$$s_{wkn}(w_{ti}|z_i) = \frac{1}{|\mathcal{R}_{w_{ti}}|} \sum_{w_r \in \mathcal{R}_{w_{ti}}} \frac{n_{w_r}^{z_i}}{n_{w_r}^*}$$

where

$\mathcal{R}_{w_{ti}} = \{w_r | (w_r, w_{ti}) \text{ is an edge in the network of concepts}\}$

$n_{w_r}^{z_i}$  is the number of times concept  $w_r$  was assigned  $t$ -topic  $z_i$  (excluding the current occurrence), and  $n_{w_r}^*$  is the number of times concept  $w_r$  was assigned any  $t$ -topic.

### 3.5 Topic assignment

To make the final assignments of  $t$ -topics to paragraphs in the test data, we use the word probabilities under both  $w$ -topics and  $t$ -topics, and the transition probability distributions for  $t$ -topics induced as a result of the iterative sampling process in the HMM framework. States correspond to  $t$ -topic assignments to paragraphs. The final assignments of  $t$ -topics to the current document is determined through a Viterbi algorithm on the HMM with the parameters described above.

## 4 Experiments

**Data** The data consists of a collection of articles about cities in English and French presented in (Chen et al., 2009).<sup>4</sup> Table 1 presents the data statistics. These documents have structure – sections, such as History, Culture, Transportation – which is what we try to recreate through this process<sup>5</sup>. The English data was manually annotated with a “clean” set of headings, to allow mapping of sections that have the same topic but slightly different headings (e.g. Culture and arts/Culture). This manual annotation process revealed 18 topics that appear in more than one document. To evaluate the cross-lingual topic alignment, the French

<sup>4</sup><http://groups.csail.mit.edu/rbg/code/mallows>.

<sup>5</sup>Yahyaoui et al. (2011)’s data did not fit this structure, and we could not use it to test this method.

paragraph labels were (manually) translated to English, if a corresponding label existed on the English side – e.g. *histoire* was translated to *history*; *voir aussi* was translated as *further reading*, its English side counterpart (instead of the more literal *see also*). Labels that had no correspondent in English were not translated.

Corpus	Language	Docs	Pars	Vocab.
CitiesEn	English	100	6670	42,603
CitiesFr	French	100	4074	31,487

Table 1: Data statistics

**Modeling parameters** The parameters of the model reflect the differences between the two languages data (e.g. English has 1.5 the number of paragraphs that French does, and also a larger vocabulary):

- for word-topic distributions: values  $< 1$  to bias towards distributions that favour high probabilities for a small set of words for both  $w$ -topics and  $t$ -topics:  $\beta = \eta = \frac{W}{100000}$  ( $W$  is the size of the vocabulary in English and French respectively);
- for topic mixtures: values  $> 1$  to produce balanced topic mixtures within paragraphs (for  $w$ -topics)  $\gamma = \frac{W}{|P|}$  ( $|P|$  is the number of paragraphs), and within documents  $\lambda = 50/K$  (for both languages, following (Griffiths and Steyvers, 2004));
- for state transitions: a high  $\kappa = 1000$  to encourage state persistence; and low  $\alpha = 0.01$  to bias the state transition model to prefer a small number of possible succeeding topics.

#### 4.1 Monolingual alignment

The performance of the model is evaluated based on the assigned  $t$ -topics on the task of cross-document alignment (Chen et al., 2009). Reference topic assignments are the section labels chosen by their authors. Following (Chen et al., 2009) we compute:

$$Rec = \frac{\sum_{h \in H} \max_{k \in K} (overlap(h, k))}{P}$$

$$Prec = \frac{\sum_{k \in K} \max_{h \in H} (overlap(h, k))}{P}$$

where  $P$  is the number of paragraphs,  $H$  is the set of section headings,  $K$  is the set of automatically assigned  $t$ -topics (themes), and  $overlap(h, k)$  is the number of paragraphs with section heading  $h$  and automatically assigned topic  $k$ .

#### 4.1.1 Settings

**Baselines** Gruber et al.’s (2007) *HTMM* models a document as a sequence of sentences. All words in a sentence have the same topic, and a binomial transition parameter determines whether the next sentence has the same topic as the current one. Chen et al. (2009) propose *GMM*, a global model for documents in a collection. They assume that the documents in the collection share the same topics, which are similarly distributed within the documents. This is captured by modeling the mixture of topics as a distribution over permutations of a topic ordering.

**Our variations** *2LDA* is the two level LDA processing as described in Section 3.2, on the texts with concept annotations. *2LDA\_HMM* adds the HMM to 2LDA to model the transition between paragraphs (Section 3.3). *2LDA\_c* adds the score based on concept relations to 2LDA. *2LDA\_c\_HMM* adds the score based on concept relations to 2LDA\_HMM.

*2LDA\_HMM no concepts* is the best performing configuration without concept annotations.

#### 4.1.2 Results

**Language models** The first result consists of the language models induced by the system. Figure 4 shows examples of the general (background), document-specific (for the English and French Wikipedia articles about the city of Montreal) and two  $t$ -topic language models. The top section of the figure shows the background and document-specific language models for the English and French data (shown side by side because the documents refer to the same cities). The background language model has captured what is usually included in a list of stopwords. This is an advantage, as stopwords may be collection-specific. The document-specific language model include words expected to be associated with the respective article titles. Sample  $t$ -topic language models are shown separately for each language.

Terms in the text are replaced with language independent concept IDs. We combine the English and French data, and apply the topic model to this data. Samples of the resulting  $t$ -topics are included in Figure 4 – columns labeled “multilingual”. The terms in *italics* represent concepts – for readability replaced with the name of the corresponding article in the English Wikipedia. We have also performed experiments with building separate topics for English and French, while shar-

Background		
en	fr	combined
the	de	the
of	la	of
and	le	and
in	et	in
to	les	de
a	des	a

Document-specific: Montreal			
en (#100)	fr (#12)	en (concepts)	fr (concepts)
montreal	motreal	<i>Montreal</i>	<i>Montreal</i>
canada	quebec	<i>Canada</i>	<i>Quebec</i>
montreals	canada	<i>Melbourne</i>	<i>Canada</i>
de	l'île	<i>Of Montreal</i>	<i>l'île</i>
french	français	<i>North America</i>	<i>français</i>
montréal	française	<i>Saint Lawrence River</i>	<i>language</i>

HL-topic #1		
en	fr	multilingual
cricket	football	<i>team</i>
basketball	tennis	<i>stadium</i>
baseball	club	<i>sport</i>
usa	clubs	<i>website</i>
tennis	coupe	<i>Olympic games</i>
hockey	stade	<i>photography</i>

HL-topic #2		
en	fr	multilingual
terminal	trains	<i>bus</i>
airlines	aéroports	<i>passenger</i>
airports	passagers	<i>train station</i>
commuter	stations	<i>rapid transit</i>
terminals	gare	<i>video</i>
cargo	route	<i>train</i>

Figure 4: Ordered sample of words from the background, document-specific and  $t$ -topic language models. The terms in italics are concepts, and we provide their English names for legibility.

ing only the language-independent concepts (i.e. after building the topics for English, we use the English topic-annotated data as observed data, and adjust the prior computed from this data while iterating over the French data). The results were very similar, and for reasons of space are not included.

The background and document-specific language models are “side-effects” of the process of cross-document and cross-language alignment. However, they can be useful. The document-specific topics could be used for (cross-language) document clustering. We took the top 20 words and concepts ranked based on tf-idf, as well as the top 20 terms in the document-specific topics for each of the 200 documents (100 for each language), and used the cosine similarity to align the documents in the two languages (i.e., to pair up the documents in English and French that were about the same city). The baseline – using tf-idf scored words – is already high (94%), due to shared city and country names. Both topic probabilities and tf-idf scored concepts perform perfectly (100%). For the task and data described in this

paper, document level alignment performed using document-specific topics can be compounded with the theme-based paragraph alignment to produce paragraph alignment between pairs of documents. We plan to study the usefulness of the document-specific topics in a more difficult environment.

**Alignment** Figure 5 presents the results in terms of precision, recall and F-score for the settings describe above to reveal the contribution of the HMM and the concept annotations: The results on the English data, for both 10 and 20  $t$ -topics, show a nice progression of the results as we add more information in the model. Adding HMM leads to better results, as it allows transition probabilities – and through this the context – to influence a  $t$ -topic assignment for each paragraph in the document. Adding concepts and their relations in the model leads to increase in recall, which is again expected since this will lead to more links between terms in different parts of the document. The results are better than previous work, with the exception of the situation when we use 20  $t$ -topics on the French data.

## 4.2 Cross-language alignment

Cross-language alignment is performed as for monolingual alignment, but we compare topic assignments across languages.

### 4.2.1 Settings

**Baselines** *Baseline 1* is based on similarity between paragraphs computed using translation probabilities from a translation table. *Baseline 2* computes paragraph alignment using a concept-based representation of each paragraph with concepts weighted according to their tf-idf score and a cosine similarity metric. Both baselines are computationally intensive because of all the pairwise comparisons between paragraphs.

**Our approach** *2LDA\_c\_HMM* We experimented with concatenating the corpora in the two languages and inducing topics from the union, or alternatively using the distributions induced on one as priors when processing the other. The results are not significantly different.

### 4.2.2 Results

Cross-lingual precision, recall and F-score computed over the entire (bilingual) collection are presented in Table 2, for 10 and 20 topics.

The high precision results for the two baselines are due in large part to singleton clusters, many of



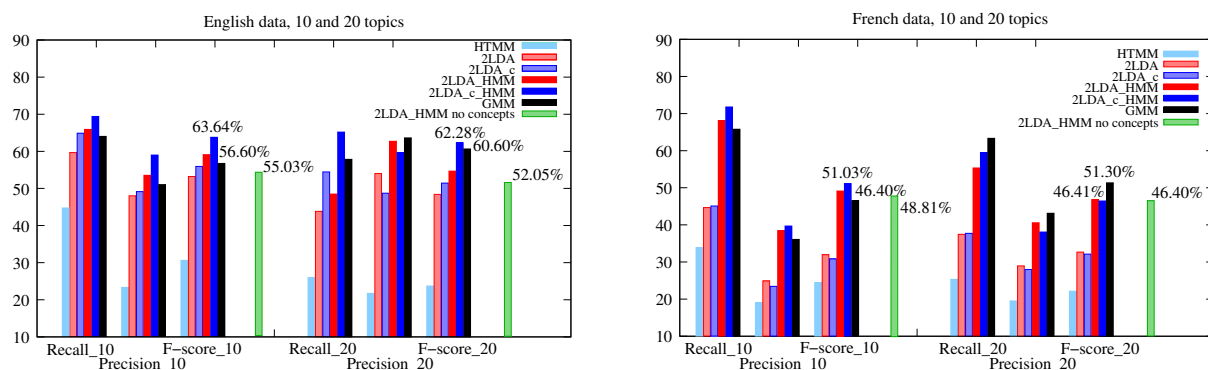


Figure 5: Alignment evaluation in terms of Precision, Recall and F-score.

Method	P	R	F
Baseline 1: transl. table	79.06	30.07	43.56
Baseline 2: concepts	81.15	29.09	47.10
2LDA_c_HMM, 10 topics	47.14	67.90	55.65
2LDA_c_HMM, 20 topics	46.85	63	53.74

Table 2: Cross-language alignment results.

which are in fact correct. Building only singleton clusters leads to an F-score of 35.13.

## 5 Conclusions

We have explored knowledge-enhanced – through concept annotations – topic models. The annotations not only help build cross-lingual topic models, but also deal with multi-word terms, polysemy and synonymy. The links between concepts are also beneficial by influencing the topic distribution and sequencing probabilities. The results are good despite the fact that we rely on an automatically built resource, and a concept identification step – itself an open research problem. We plan to investigate a joint approach to topic modeling and concept identification.

We have augmented the traditional HMM model with language models within a paragraph, and a Dirichlet distribution at the level of state (high-level topic) transitions. Modeling a text on two levels gives a layered view, which can be used to structure the commonly used bag-of-word representation of texts. Separating words that contribute to the topic segmentation from those that pertain to a background language or document-specific model helps the system focus on the most relevant terms for segmentation. This could become an advantage when aligning corpora in different languages – one need only establish parallelism between topic-specific terms.

## References

- [Barzilay and Elhadad2003] Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, 11–12 July 2003, pages 25–32.
- [Beal et al.2002] Matthew Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. 2002. The infinite hidden Markov model. In *NIPS*, pages 577–584.
- [Blei and Moreno2001] David M. Blei and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *SIGIR*, pages 343–348.
- [Boyd-Graber and Blei2009] Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Uncertainty in Artificial Intelligence*.
- [Boyd-Graber et al.2007] Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, Prague, Czech Republic, 28–30 June 2007.
- [Chen et al.2009] Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Content modeling using latent permutations. *JAIR*, (36).
- [Daumé III and Marcu2006] Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17–21 July 2006, pages 305–312.
- [Eisenstein and Barzilay2008] Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 334–343.



- [Fahrni et al.2011] Angela Fahrni, Vivi Nastase, and Michael Strube. 2011. HITS' graph-based system at the NTCIR-9 cross-lingual link discovery task. In *Proceedings of the 9th NTCIR Workshop Meeting*, Tokyo, Japan, 6-9 December 2011, pages 473–480.
- [Fox et al.2010] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S Willsky. 2010. A sticky HDP-HMM with application to speaker diarization.
- [Fung and McKeown1997] Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of 5th International Workshop of Very Large Corpora (WVLC-5)*, Hongkong.
- [Griffiths and Steyvers2004] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.
- [Gruber et al.2007] Amit Gruber, Michael Rosen-Zvi, and Yair Weiss. 2007. Hidden topic Markov models. In *AISTATS*.
- [Gupta et al.2013] Rajdeep Gupta, Santanu Pal, and Sivaji Bandyopadhyay. 2013. Improving mt system using extracted parallel fragments of text from comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 69–76, Sofia, Bulgaria.
- [Jagarlamudi and Daumé III2010] Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010*, pages 444–456.
- [Mimno et al.2009] David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 6–7 August 2009, pages 880–889.
- [Mulbregt et al.1998] P. Van Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron. 1998. Text segmentation and topic tracking on broadcast news via a hidden markov model approach. In *ICSLP-98*, pages 2519–2522.
- [Nastase and Strube2013] Vivi Nastase and Michael Strube. 2013. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.
- [Ni et al.2009] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain, April 20-24, 2009, pages 1155–1156.
- [Paul and Girju2009] Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *EMNLP*, pages 1408–1417.
- [Teh et al.2003] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2003. Hierarchical dirichlet processes. *JASA*, 101.
- [Tiedemann2011] Jörg Tiedemann. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- [Titov and McDonald2008] Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *WWW*.
- [Vulić et al.2011] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Cross-language information retrieval with latent topic models trained on a comparable corpus. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology*, AIRS'11, pages 37–48.
- [Wang et al.2011] Hongning Wang, Duo Zhang, and ChengXiang Zhai. 2011. Structural topic model for latent topical structure analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., USA, 19–24 June 2011, pages 1526–1535.
- [Yahyaee et al.2011] Sirvan Yahyaee, Marco Bonzanini, and Thomas Roelleke. 2011. Cross-lingual text fragment alignment using divergence from randomness. In *Proceedings of the 18th International Conference on String Processing and Information Retrieval*, SPIRE'11, pages 14–25.
- [Zhai et al.2004] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *KDD*.
- [Zhang et al.2010] Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pages 1128–1137.